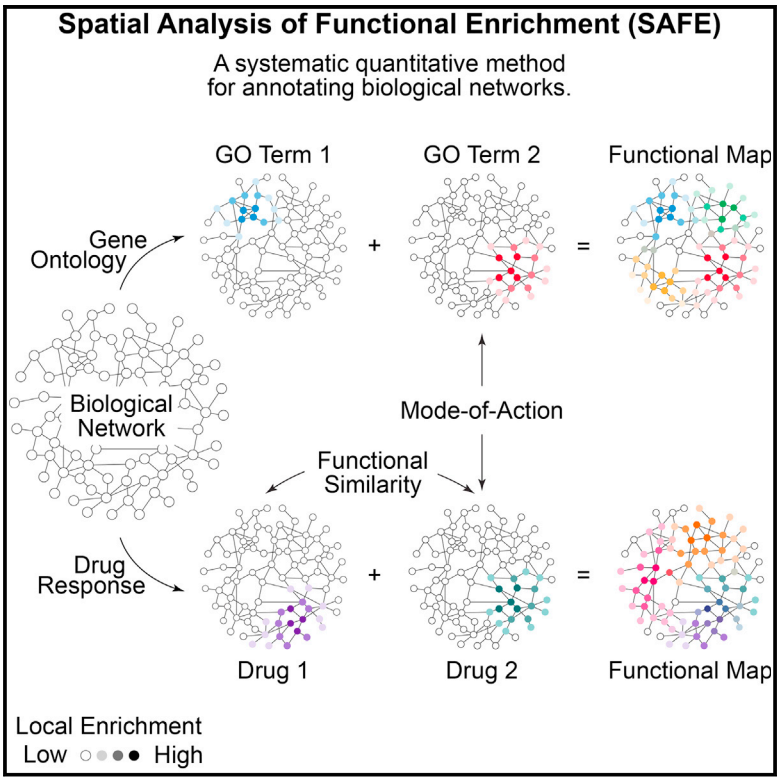
Report

Cell Systems

Systematic Functional Annotation and Visualization of Biological Networks

Graphical Abstract



Highlights

- SAFE stands for spatial analysis of functional enrichment
- SAFE annotates networks by mapping local enrichment for functions and phenotypes
- SAFE's outcome is intuitive, quantitative, robust to noise, and sensitive to signal
- SAFE uncovers new biological mechanisms, including resistance to chemotherapy

Authors

Anastasia Baryshnikova

Correspondence

abarysh@princeton.edu

In Brief

Spatial analysis of functional enrichment (SAFE) is a systematic, quantitative method for mapping local enrichment for functional attributes in biological networks. Using the yeast genetic interaction network as a test case, SAFE proved to be accurate, robust, and predictive of new biological mechanisms, such as resistance to the anti-cancer drug bortezomib.

Baryshnikova, 2016, Cell Systems *2*, 412–421
June 22, 2016
http://dx.doi.org/10.1016/j.cels.2016.04.014

OPEN ACCESS
CellPress

Cell Systems
Report

Systematic Functional Annotation and Visualization of Biological Networks

Anastasia Baryshnikova[1,*]
[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
*Correspondence: abarysh@princeton.edu
http://dx.doi.org/10.1016/j.cels.2016.04.014

SUMMARY

Large-scale biological networks represent relationships between genes, but our understanding of how networks are functionally organized is limited. Here, I describe spatial analysis of functional enrichment (SAFE), a systematic method for annotating biological networks and examining their functional organization. SAFE visualizes the network in 2D space and measures the continuous distribution of functional enrichment across local neighborhoods, producing a list of the associated functions and a map of their relative positioning. I applied SAFE to annotate the *Saccharomyces cerevisiae* genetic interaction similarity network and protein-protein interaction network with gene ontology terms. SAFE annotations of the genetic network matched manually derived annotations, while taking less than 1% of the time, and proved robust to noise and sensitive to biological signal. Integration of genetic interaction and chemical genomics data using SAFE revealed a link between vesicle-mediate transport and resistance to the anti-cancer drug bortezomib. These results demonstrate the utility of SAFE for examining biological networks and understanding their functional organization.

INTRODUCTION

Understanding the functional organization of living cells is essential for predicting cell behavior in normal and diseased conditions and designing effective therapeutic strategies to control it. Budding yeast *Saccharomyces cerevisiae* is particularly useful for elucidating the organization of a cellular system, owing to the availability of extensive molecular interaction networks that map physical, biochemical, and phenotypic relationships between nearly all genes in the genome (Gavin et al., 2006; Krogan et al., 2006; Zhu et al., 2007; Tarassov et al., 2008; Yu et al., 2008; Mo et al., 2009; Costanzo et al., 2010). However, functional annotation of these networks—that is, determining which biological functions are represented in each network, which parts of the network these functions are associated with, and how they are related to one another—is challenging, especially due to the scarcity of rigorous statistical methods and reproducible workflows.

Systematic annotation of a biological network can be formulated as a three step process. First, obtain a comprehensive map of the network showing all of its nodes and their connections to one another. This map can be produced using a network layout algorithm that embeds the network in 2D or 3D space and positions all nodes based on their connectivity (Kobourov, 2012). Second, gather multiple independent datasets of functional information that characterize all nodes relative to one another based on a variety of parameters (e.g., cellular localization, transcriptional response to a perturbation, and mutant phenotype). Such functional resources are readily available in yeast, thanks to the development of numerous genomic assays and a long history of literature curation (Botstein and Fink, 2011). Finally, implement automated statistical procedures to overlay functional data onto the network map and identify functionally coherent regions. Functional regions have been discovered on a case-by-case basis (Khatri et al., 2012; Mitra et al., 2013), but no method is currently equipped to identify them exhaustively, locate them relative to one another, and create a functional map of the network that is comprehensive, quantitative, and intuitive to biologists.

In principle, some existing approaches could be repurposed to produce such a functional map. However, these approaches would require extensive modifications and/or additional post-processing steps that would drastically reduce their scalability for automation. For example, several network algorithms have been developed to search for sets of interconnected nodes that share a common phenotype or a consistent response across experimental conditions (Mitra et al., 2013). The main purpose of these algorithms is to evaluate experimental datasets and identify the most promising candidate genes supported by network connectivity. Since the network itself is not the focus of the analysis, but is only independent supporting evidence, these methods cannot be directly applied for comprehensive annotation tasks. Similarly, network clustering algorithms could be potentially used to identify sets of densely connected nodes that correspond to known, as well as novel, functional modules (Newman, 2006). However, clustering disregards loosely connected nodes, causing many sparse, yet functionally coherent, network regions to remain unnoticed. In addition, clustering partitions the network into discrete and, in some cases, overlapping subnetworks, which must be separately annotated and integrated back together to provide a global functional view of the network. Since rapid and reproducible integration of functional annotations has yet to be achieved systematically, the use of clustering algorithms for annotating biological networks is currently impractical. More recently, several methods have been proposed to identify gene attributes that co-cluster within

a network more closely than expected by random chance (Cornish and Markowetz, 2014; Menche et al., 2015). While significant co-clustering does indicate that an attribute is strongly associated with the network, it does not reveal in which part of the network the association occurs and how it relates to other attributes. As a result, both coverage and specificity of attribute associations cannot be assessed and a true functional map of the network cannot be built.

Here, to address the limitations of the existing methods, I describe spatial analysis of functional enrichment (SAFE), an automated procedure for annotating biological networks and generating quantitative and intuitive functional maps. By mapping a function (or a phenotype) to a specific part of the network, SAFE provides statistical evidence, as well as an intuitive visual representation, for the positioning of the function within the network and thus facilitates the investigation of the network global functional organization.

RESULTS

Overview of SAFE

SAFE annotates a biological network by calculating and displaying local enrichment for a set of functional attributes (Figure 1; Experimental Procedures). SAFE first generates a 2D map of the network by applying a force-directed network layout algorithm or importing a network map produced by a third-party software, such as Cytoscape (Shannon et al., 2003) (Figure 1A). In a network map, nodes are positioned based on an equilibrium of forces that reflects network topology: e.g., connected nodes attract each other, whereas disconnected nodes are repelled (Kobourov, 2012).

For every node on the network map, SAFE defines a local neighborhood, i.e., a set of nodes located within a certain distance, but not necessarily directly connected (Figure 1B). While several distance metrics are available, the default option is the map-weighted shortest path length (MSPL), whereby edges are weighted according to their physical length on the map and distance between two nodes is defined by the shortest physical path connecting them (Figure 1B).

For every neighborhood, SAFE calculates a set of quantitative scores, each corresponding to the sum of the neighbors' values for a functional attribute (Figure 1C). Attributes may be binary, such as annotations to gene ontology (GO) terms, or quantitative, such as phenotypic response to a perturbation. To estimate the significance of a neighborhood's score for an attribute, SAFE computes a hypergeometric p value (for binary annotations) or an empirical p value (for quantitative annotations) based on 1,000 network randomizations that preserve network topology but reshuffle attribute assignments (Figure 1C). A log enrichment score ($-\log_{10} p$, where p is the enrichment p value, corrected for multiple testing across all attributes) is then assigned to the center of the neighborhood (Figure 1D).

The enrichment scores for any given attribute across all nodes on the map define the attribute's enrichment landscape (Figure 1E), which has a characteristic size (i.e., the number of enriched neighborhoods), shape (i.e., the map area covered by the enriched neighborhoods) and relief (i.e., the peaks and valleys in the enrichment). Through these properties, the enrichment landscape reflects the distribution of the attribute across the map and measures the strength of its association with the network.

To combine multiple enrichment landscapes into a single functional map of the network, SAFE pairs each attribute with a unique color and sums the colors in each neighborhood proportionally to the enrichment of their corresponding attributes (Figure 1F). As a result, pure colors indicate regions dominated by a single functional attribute, whereas hybrid colors represent multi-functional regions (Figure 1F). This approach produces a quantitative representation of functional enrichments throughout the network and a comprehensive view of their network localization.

When several attributes are closely related and map to the same region of the network (as occurs, for example, for a GO term and its descendants), SAFE can combine them into a single functional domain using the similarity of their enrichment landscapes. To facilitate the interpretation of the resultant functional map, all attributes within a domain are assigned the same color and labeled by a single tag list, composed of the five most recurrent words among the attributes' names.

Benchmarking SAFE on the Yeast Genetic Interaction Similarity Network

To test the SAFE method, I applied it to annotate a yeast genetic interaction similarity (GIS) network using GO biological process terms as attributes and compared the results to a manual annotation previously published (Costanzo et al., 2010) (Figure 2; Experimental Procedures). A genetic interaction is a phenotypic relationship between two genes that occurs when the phenotype of the double mutant deviates from the expected combination of the phenotypes of the two single mutants (Baryshnikova et al., 2013). Genes sharing similar genetic interactions often share a common biological function (Baryshnikova et al., 2010b) and form a functional network that connects most genes in the yeast genome (Costanzo et al., 2010). A high-confidence version of this network, consisting of 2,838 nodes and 10,016 edges (Figure 2A), was annotated in its original study via a thorough, yet mostly manual, procedure and therefore provides a good test case for SAFE.

Using a map of the GIS network produced by the spring-embedded network layout in Cytoscape, SAFE measured local enrichment for 4,373 GO biological process terms, each associated with at least one yeast gene, and revealed a great variation in the size, shape, and relief of GO term enrichment landscapes (Figures 2B–2D and S1A). The vast majority of GO terms (84%) were only enriched within the neighborhoods of ten or fewer genes, indicating that these terms were too small or too sparsely distributed throughout the network to be informative about its functional organization (Figure 2B). The remaining GO terms were enriched in more than ten neighborhoods, but varied in the spatial distribution of their enrichments: 12% of GO terms (506 of 4,373) were region specific as they displayed a single peak of enrichment in a single region of the network (Figure 2C), whereas 4% of GO terms (174 of 4,373) were multi-regional with two or more peaks in different network regions (Figure 2D). The presence of multiple peaks suggested that each multi-regional term was comprised of several subgroups of genes that localize separately in the

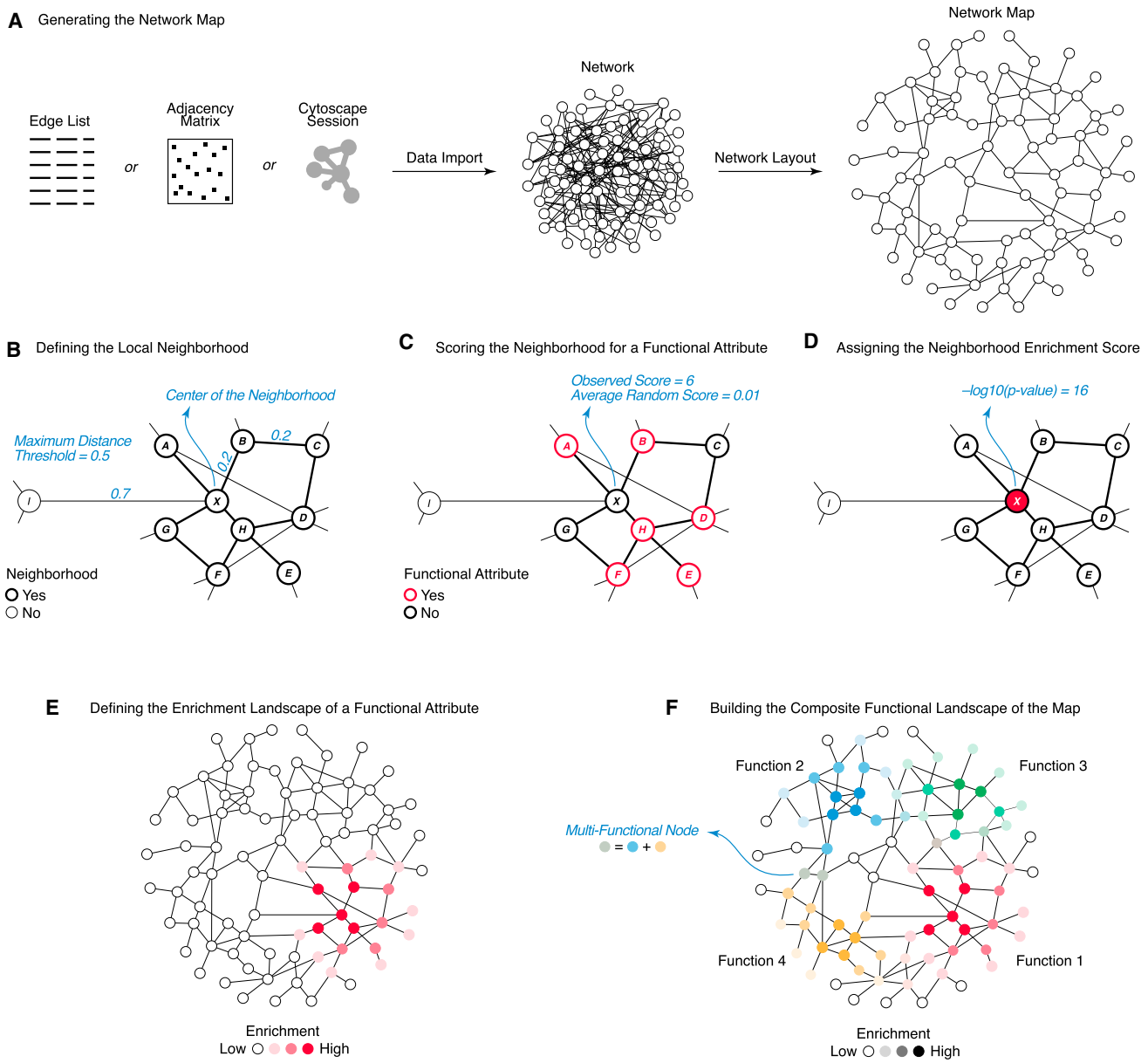


Figure 1. SAFE

The source code and a standalone application are provided as http://dx.doi.org/10.17632/wdxwy8gmrz.1.

(A) Given a biological network, as an edge list or an adjacency matrix, SAFE generates a 2D map of the network by applying a force-directed network layout algorithm. Alternatively, SAFE can import a network map directly from the network visualization software Cytoscape (Shannon et al., 2003).

(B) For each node *X* on the network map, SAFE defines the local neighborhood of *X* by identifying all other nodes (*A*-*H*) located closer than a maximum distance threshold. By default, node distance is measured using the map-weighted shortest path length (MSPL); however, other distance measures are also available.

(C) For each neighborhood, SAFE sums the neighbors' values for a functional attribute of interest, compares the result to random expectation, and computes a significance p value.

(D) The p value, corrected for multiple testing across all attributes, is log-transformed into a neighborhood enrichment score and assigned to node *X*.

(E) The neighborhood enrichment scores for any given attribute across all nodes in the network define the attribute's enrichment landscape. Color intensity represents variation in enrichment scores.

(F) Attributes are assigned unique colors and summed proportionally to their relative enrichment within each neighborhood. As a result, pure colors indicate regions with a single predominant function, whereas hybrid colors correspond to multi-functional regions. Enrichment scores, color assignments, and node-attribute associations are provided as text output.

network and thus may be functionally distinct. Notably, each subgroup appeared to be well covered by at least one region-specific landscape, suggesting that region-specific GO terms were sufficient to annotate the entire network. As a result, sparse and multi-regional terms were excluded from further analysis (Discussion).

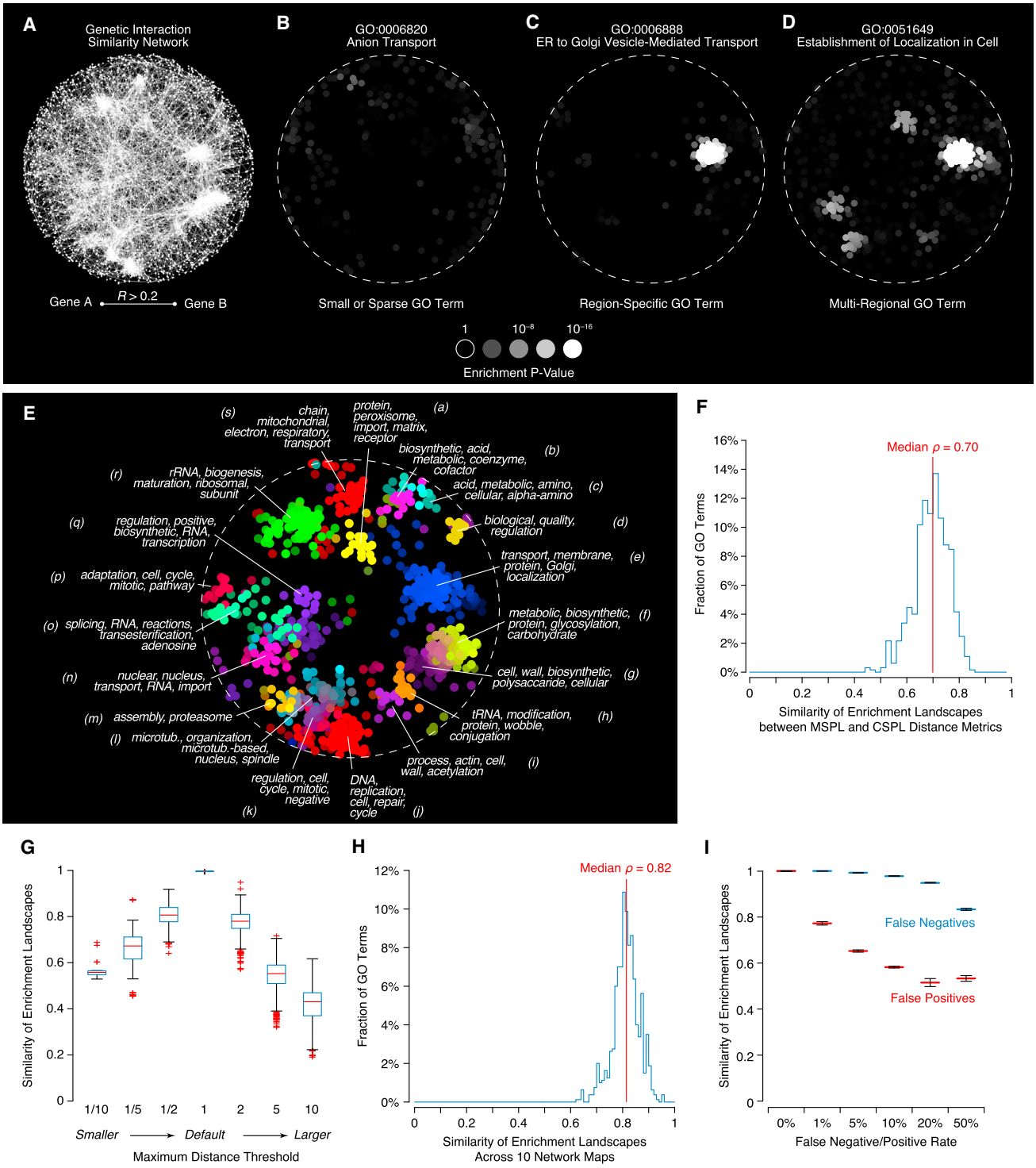


Figure 2. SAFE Annotation and Benchmarking of the Yeast GIS Network with GO Biological Process Terms

(A) A map of the network, containing 2,838 nodes and 10,016 edges, was originally constructed in Costanzo et al. (2010). Two genes (nodes) were connected if their genetic interaction profiles showed a Pearson correlation coefficient (*R*) greater than 0.2. Nodes were organized in 2D space by applying the spring-embedded layout algorithm in Cytoscape.

(B–D) SAFE annotation of the network with 4,373 GO biological process terms showed that different GO terms have different types of enrichment landscapes. Default SAFE parameters were applied. (B) Most GO terms (84%), including "anion transport" (GO: 0006820), were poorly enriched in the network. (C) 12% of GO terms, including "ER to Golgi vesicle-mediated transport" (GO: 0006888), showed a region-specific enrichment landscape. (D) 4% of GO terms, including "establishment of localization in cell" (GO: 0051649), showed multi-regional enrichment landscapes.

(legend continued on next page)

Since many of the 506 region-specific GO terms mapped to the same network region, their contributions toward annotating the network were partially redundant. To minimize redundancy and simplify the annotation process, SAFE grouped the terms into functional domains based on the similarity of their enrichment landscapes. The resultant 19 domains, represented by different colors and labeled with tag lists, formed a comprehensive, systematic, and quantitative GO-based functional annotation map of the GIS network (Figure 2E; Data S1).

The functional map, produced by SAFE, was highly consistent with the manual annotation, carried out in (Costanzo et al., 2010). Specifically, SAFE identified all of the manually annotated regions (p values $< 2 \times 10^{-4}$, Fisher's exact test) and associated them with GO terms that matched their manually assigned labels (Figure S1B). Notably, SAFE also recognized three previously unannotated network regions that were missed, likely because of their reduced size and specific localization (Figures 2E and S1B).

Analysis of Robustness

Next, I sought to assess the robustness of SAFE to variations in distance metrics, neighborhood size, network layout, and noise in annotations. In annotating the GIS network, SAFE used default parameters and assumed that (1) the map-weighted shortest path length (MSPL) is a reliable measure of node distance, (2) only nodes within a certain distance are part of a neighborhood, and (3) neighborhoods remain unchanged across multiple runs of the spring-embedded network layout algorithm, which, due to an initial randomization step, is non-deterministic.

To validate the first assumption, I tested how much neighborhood enrichment would change if node distances were defined by a map-independent metric, such as, for example, the correlation-weighted shortest path length (CSPL). In CSPL, each edge is weighted by $1–R$, where R is the correlation of the genetic interaction profiles of the two connected genes and a value that has not been used in constructing the map (Figure S2A; Experimental Procedures). Systematic comparison of the enrichment landscapes produced by MSPL and CSPL showed that their median similarity is $\rho = 0.70$ (Spearman's rank correlation; Figure 2F). This indicates that, regardless of the distance metric, most neighborhoods remain enriched for the same GO terms to a similar degree and, thus, MSPL is reliable at estimating node distances.

By default, only nodes whose distance is in the lowest 0.5^{th} percentile of all network distances are in each other's neighborhoods. I tested how much this assumption affects neighborhood enrichment by examining a wide range of distance thresholds. I found that increasing or decreasing the default threshold up to 2-fold produces highly similar enrichment landscapes (median $\rho = 0.78$ and $\rho = 0.81$, respectively; Figure 2G) and has a limited impact on the set of enriched GO terms (Figure S2B). This indicates that neighborhood enrichment is fairly insensitive to neighborhood size, and the choice of a distance threshold within a 2-fold range is not critical.

Finally, I assessed the robustness of neighborhood enrichment to the non-deterministic nature of the spring-embedded layout algorithm, which moves all nodes to new equilibrium positions at every run (Experimental Procedures). By repeatedly applying the layout and comparing the enrichment landscapes of the resultant network maps, I found that, on average, the enrichment landscapes of any two independent maps were highly similar (median across all GO terms $\rho = 0.82$; Figures 2H and S2C–S2E). This indicates that, despite differences in absolute node positions across layout runs, the neighborhoods remain largely unchanged. To eliminate the residual variability and ensure complete reproducibility, SAFE can be set to control the randomization step of the layout algorithm and produce identical node positions at every run (Experimental Procedures).

The accuracy of network annotation may also depend on the quality of the functional annotation standard. To test the robustness of SAFE to annotation noise, I systematically altered all GO biological process terms by randomly introducing varying amounts of false-positive or false-negative annotations and compared the resultant enrichment landscapes to their error-free versions (Figure 2I). I found that a false-negative rate (FNR) as high as 20% failed to affect nearly any of the enrichment landscapes (median $\rho = 0.95$), whereas a false-positive rate (FPR) as low as 1% had a much larger impact (median $\rho = 0.78$). Importantly, while both false-positive and false-negative annotations decreased the number of GO terms enriched within the neighborhoods of at least ten genes (by 20% and 27% at FNR = 20% and FPR = 1%, respectively), neither type of error gave rise to new, previously unobserved GO term enrichments (Figures S2F and S2G).

SAFE Facilitates the Integration of Functional Datasets

Because of the inherent biases and limitations of all functional standards, annotating a network with a single type of biological information, such as GO, is unlikely to provide a full picture of the network's functional organization. A more valid strategy is to use multiple independent sources of functional data and apply them iteratively to annotate the same network. Such an approach would not only provide a more realistic description of the network but could also reveal unexpected relationships between data types.

In yeast, chemical genomics provides a rich source of functional information (Ho et al., 2011). In a chemical genomic screen, a genome-wide collection of yeast mutants is grown in the presence of a chemical compound, and the relative fitness of each

(E) SAFE constructed a functional map of the network by combining all region-specific GO terms into 19 functional domains (*a*-*s*) based on the similarity of their enrichment landscapes. Different colors represent different functional domains. Each domain is labeled with a tag list, composed of the five words that occur most frequently within the names of the associated GO terms. The complete list of GO terms is reported in Data S1.
(F–I) Comparative analysis of GO term enrichment landscapes across several variations in input data and settings. In all cases, only landscapes with at least ten enriched neighborhoods were considered. (F) Distribution of Spearman's rank correlations between enrichment landscapes computed using the MSPL and the CSPL distance metrics. (G) Distribution (median, first to third quartiles ± 1.5 × range between the quartiles) of Spearman's rank correlations between enrichment landscapes across a range of maximum distance thresholds. (H) Distribution of Spearman's rank correlations between enrichment landscapes across ten different network maps. (I) Average Spearman's rank correlations between enrichment landscapes before and after the introduction of false-negative (blue) or false-positive (red) annotations. Each alteration was performed 1,000 times. Error bars represent standard deviations of 1,000 medians.
See also Figures S1 and S2.

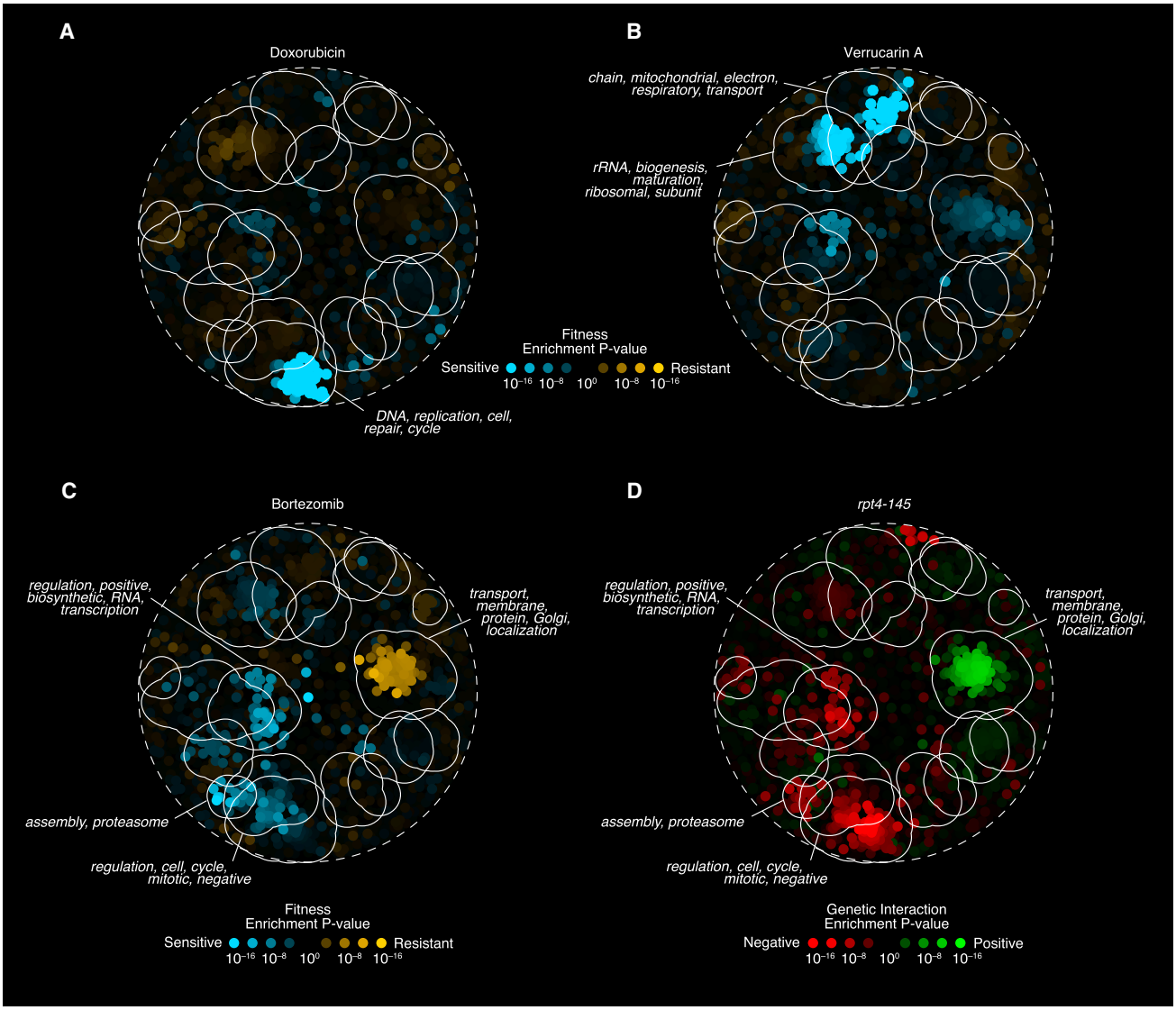


Figure 3. SAFE Annotation of the Yeast GIS Network with Chemical Genomic Data

For reference, the outlines of the 19 GO-based functional domains (Figure 2E) are shown.

(A) Fitness enrichment landscape of doxorubicin, an inhibitor of DNA replication.

(B) Fitness enrichment landscape of verrucarin A, an inhibitor of protein synthesis and mitochondrial function.

(C) Fitness enrichment landscape of bortezomib, a proteasome inhibitor.

(D) Genetic interaction enrichment landscape of *rpt4-145*, a temperature-sensitive allele of an essential proteasome subunit.

See also Figure S3.

mutant is measured with respect to an untreated control (Giaever et al., 2002). Identifying mutants that are particularly resistant or sensitive to a given compound is instrumental for mapping pathways that mediate the compound's toxicity or are required to protect the cell against its detrimental effects (Ho et al., 2011). I hypothesized that SAFE could assist the identification of these pathways by annotating the GIS network with chemical genomic data and identifying functional network regions enriched for sensitivity or resistance to chemicals.

To test this hypothesis, I used a recent chemical genomics dataset that measured quantitative fitness scores for ~5,000 yeast homozygous deletion mutants upon exposure to 132 chemical compounds with known modes of action (Hoepfner et al., 2014). Using these data, SAFE generated 132 compound-specific fitness enrichment landscapes that mapped the relative distribution of sensitive and resistant mutants throughout the GIS network (Figures 3A–3C).

The analysis of the fitness enrichment landscapes in the context of the GO biological process map showed that the landscapes are highly consistent with our current knowledge about the compounds' modes of action (Figures S3A and S3B; Data S1). For example, mutants sensitive to doxorubicin, a DNA intercalator that prevents DNA replication by blocking the progression of topoisomerase II (Tacar et al., 2013), were specifically

enriched in the network region associated with DNA replication and repair GO terms (Figure 3A). Similarly, regions enriched for ribosome- and mitochondria-related GO terms were also overrepresented for mutants sensitive to verrucarin A, a protein synthesis inhibitor (Hernández and Cannon, 1982) with reported toxicity toward mitochondria (Schappert and Khachatourians, 1986) (Figure 3B).

Case Study: SAFE Reveals a Mechanism of Resistance to Bortezomib

In addition to known compound modes of action, SAFE can also uncover new response patterns. An example is the fitness enrichment landscape of bortezomib, a proteasome inhibitor approved for treating multiple myeloma and mantle cell lymphoma and undergoing clinical trials for several other types of cancer (Chen et al., 2011) (Figure 3C). SAFE analysis showed that, in yeast, sensitivity to bortezomib was specifically overrepresented in network regions associated with proteasome-mediated protein degradation, cell-cycle control, and transcriptional regulation (Figure 3C). These results are consistent with observations in human cells suggesting that bortezomib indirectly promotes programmed cell death by preventing the degradation of pro-apoptotic factors (Chen et al., 2011) and is synergistic with histone deacetylase inhibitors (Yu et al., 2003), which regulate transcription.

Notably, however, SAFE also showed that resistance to bortezomib was strongly linked to the network region enriched for secretion and vesicle-mediated transport GO terms (Figure 3C). Although several reports have suggested that proteasome inactivation may cause endoplasmic reticulum (ER) stress, due to an accumulation of misfolded proteins in the ER (Lee et al., 2003; Obeng et al., 2006), it has not been anticipated that loss-of-function mutations in ER- or other vesicle-related functions could alleviate this or other proteasome-related stresses. A direct examination of bortezomib's fitness data confirmed SAFE's enrichment analysis: the four mutant strains with the highest resistance to the drug carried complete or partial deletions in *YTP6*, *RIC1*, or *RGP1*, genes regulating the formation, motility, and fusion of vesicles traveling to and from the Golgi compartment (Siniossoglou et al., 2000). In addition, several other proteins involved in Golgi-related transport were among the top 15 most resistant mutants.

The statistical significance of this finding was also supported by a network-independent gene set enrichment analysis (GSEA) of bortezomib's fitness data. GSEA determines whether members of a functional group tend to occur at the top or at the bottom of a ranked gene list and measures the probability of such distribution to arise by random chance (Subramanian et al., 2005). By applying GSEA to the ranked list of bortezomib's fitness scores and all 4,373 GO biological process terms, I confirmed that processes such as "intra-Golgi vesicle-mediated transport" (GO: 0006891) and "cytoplasm to vacuole targeting (CVT) pathway" (GO: 0032258) are significantly associated with the most highly resistant mutants ($p < 0.001$ and $p = 0.001$, respectively; FDR < 0.05; Data S2). Compared to the results produced by SAFE, however, the primacy of these pathways in resisting bortezomib was much less apparent in the GSEA results: the majority (58%) of the 48 significant GO terms detected by GSEA involved ion homeostasis, regulation of intracellular pH, and other distantly related functions associated with only moderately resistant mutants (Data S2). Such a discrepancy suggests that, by leveraging network topology, SAFE can detect functional signals not readily apparent in network-independent analyses.

To further validate the connection between vesicle-mediate transport and resistance to bortezomib, I asked whether, similarly to other drugs (Parsons et al., 2004; Costanzo et al., 2010), the effect of bortezomib could be mimicked by the genetic inactivation of its molecular target, i.e., the proteasome. Specifically, I hypothesized that mutations providing a relative growth advantage in the presence of bortezomib should also display a relatively high fitness when combined with proteasome mutants and should therefore result in positive genetic interactions.

To test this hypothesis, I obtained quantitative negative and positive genetic interactions for members of the core and the regulatory particles of the yeast proteasome (Experimental Procedures; Data S2) and used them to annotate the GIS network with SAFE. For every gene encoding a proteasomal subunit, SAFE calculated a genetic interaction enrichment landscape that mapped local enrichment for the gene's negative and positive interactors throughout the network (Figure 3D). I found that at least 7 of the 13 tested subunits of the proteasome regulatory particle were significantly enriched for positive genetic interactions in the vesicle transport region (Figure S3C). In particular, *rpt4-145*, a temperature-sensitive mutant in an essential ATPase of the proteasome regulatory particle that preferentially contributes to ER-associated protein degradation (ERAD) (Lipson et al., 2008), showed a genetic interaction enrichment landscape remarkably similar to bortezomib's fitness enrichment landscape (Figure 3D). It is important to note that a direct comparison between the fitness profiles of *rpt4-145* and bortezomib would also have revealed their similarity ($r = 0.22$, $p < 10^{-31}$); however, determining what drives this similarity, i.e., specifically which positive and negative interactions are common to both profiles, could not have been accomplished by correlation analysis alone.

These findings support the hypothesis that mutations in secretory functions may partially compensate for proteasome inactivity and alleviate the damaging effects of bortezomib treatment, although the precise mechanisms of this alleviation are still unknown. Considering that resistance to bortezomib is a common complication in treating multiple myeloma and other cancers (Murray et al., 2014), understanding its molecular mechanisms is critical for designing effective drug combinations and personalized therapies. The identification of secretory pathways as a potential focus of these new therapies illustrates the power of SAFE to uncover novel biological responses through comprehensive annotation of biological networks with multiple independent functional standards.

Annotation of More Complex Networks, Including Protein-Protein Interactions

Compared to other biological networks, the GIS network is relatively sparse and modular (Figure 2A) and thus may be particularly amenable to annotation. To assess whether SAFE can also annotate more complex networks, I first tested whether it can detect functional enrichment in a denser version of the GIS network, obtained by decreasing the minimum GIS threshold required for connectivity (Figure S4A; Experimental Procedures). Annotation of a GIS network with 40%–240% more edges than

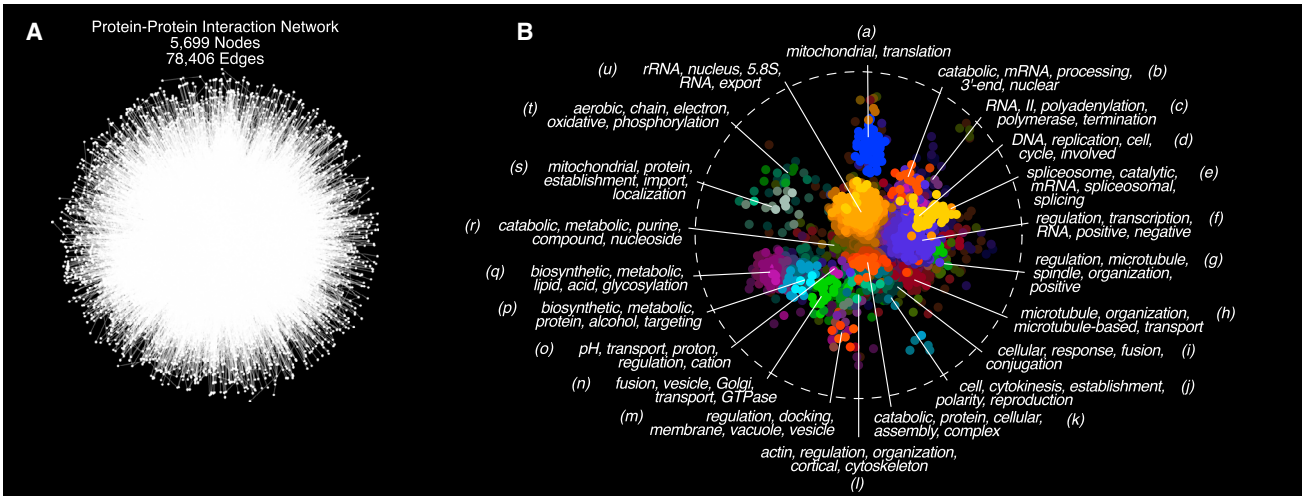


Figure 4. SAFE Annotation of the Yeast Protein-Protein Interaction Network

(A) A protein-protein interaction network was constructed by connecting all yeast protein pairs that share a physical interaction type in the BioGRID database. A map of the network was generated by applying the spring-embedded layout algorithm in Cytoscape.

(B) The map was annotated with SAFE using 4,373 GO biological process terms. All region-specific GO terms were combined into 21 functional domains (*a*-*u*) based on the similarity of their enrichment landscapes. Different colors represent different functional domains. Each domain is labeled with a tag list, composed of the five words that occur most frequently within the names of the associated GO terms.

See also Figure S4.

the original produced similar GO term enrichment landscapes (median ρ = 0.61–0.75; Figures S4A and S4B), suggesting that SAFE is sensitive to coherent functional structure, regardless of network density.

It is also possible that more complex biological networks do not have a functional structure as prominent as GIS and thus are more difficult to annotate. Therefore, I examined the extensive protein-protein interaction network that maps 78,406 physical binding events among 5,699 yeast proteins (Experimental Procedures). While protein-protein interactions are known to preferentially connect members of the same protein complex and other functionally related proteins (Cusick et al., 2005), the map of the network showed no visible topological structure on a global scale (Figure 4A). However, SAFE revealed 21 large functional domains in the network, and each domain was associated with a distinctive GO term enrichment landscape and a unique list of enriched GO terms (Figure 4B; Data S3). The resultant functional map strongly suggests that physical binding, just like GIS, assembles proteins into large functional communities that extend beyond protein complexes and molecular pathways. The relative network positioning of these communities, alongside their GIS counterparts, may provide new insight into the global functional organization of the yeast cell.

DISCUSSION

Here, I describe the development, validation, and application of SAFE, an automated method for annotating biological networks and examining their functional organization. Given a network and a visual map of its connectivity, SAFE localizes all network regions enriched for one or more functional attributes, such as GO terms or quantitative phenotypes. In doing so, SAFE answers three fundamental questions. First, are any regions of this network specifically associated with a given function or phenotype? Second, where in the network are these regions localized? And third, how does their localization compare to that of other functions or phenotypes? By answering these questions, SAFE builds a functional map of the network and enables the investigation of inter-process relationships within the cell.

Through the power of visualization, SAFE could also be used to improve our understanding of functional standards. For example, SAFE annotation of the GIS network with GO biological process terms showed that some GO terms are enriched in a single network region, whereas others are multi-regional (Figures 2B–2D). Although multi-regional GO terms tended to be generally larger than region-specific terms (p value = 10^{-49}, rank-sum test), their size distributions were largely overlapping (Figure S1C), suggesting that term size is not solely responsible for landscape differences. An interesting possibility is that region-specific GO terms share a similar level of functional specificity, which is defined by the topology of the GIS network. If that is the case, these terms could be used to delineate a cross-section of the GO hierarchy and create a flat subset of GO annotations, analogous to GO slim (Ashburner et al., 2000; http://geneontology.org/page/go-slim-and-subset-guide). Flat annotation resources are instrumental in genomics, thanks to their smaller size and lower redundancy (Myers et al., 2006; Costanzo et al., 2010). SAFE may offer the possibility to generate data-driven network-specific flat annotation standards that would enable more targeted functional analyses and simplify their interpretation.

The interpretation of SAFE functional maps would greatly benefit from a better understanding of network layout algorithms. Data-driven network layouts, such as spring embedding (http://wiki.cytoscape.org/Cytoscape_3/UserManual - Cytoscape_3.2BAC8-UserManual.2BAC8-Navigation_Layout.Automatic_Layout_Algorithms), are unsupervised methods for

organizing nodes based on their connectivity and are de facto multi-dimensionality reduction procedures. In its default setup, SAFE relies on layouts to identify local neighborhoods and map their functional enrichment. However, very little is known about how a particular layout should be chosen. Despite their great potential for uncovering hidden patterns within the data, layouts are typically used to generate esthetically pleasing network visualizations and are rarely the basis of any systematic network analysis. As a result, we have limited experience in evaluating network layouts and a poor understanding of their relative performance in the context of different networks. SAFE may be useful for quantitatively evaluating alternative layouts for the same network using a common set of functional attributes. Ideally, such analyses could identify optimal layouts for every network type and establish common ground for comparing networks.

Quantitative comparison of biological networks is a major goal of systems biology (Sharan and Ideker, 2006). A deep understanding of how genes, pathways, and processes are connected across different network types will aid in developing successful strategies for integrating networks into a single comprehensive model of a living cell. By mapping enrichment for the same functional attributes in different networks (Figures 2 and 4), SAFE can make an important contribution to this goal. However, rigorous statistical approaches must be implemented to compare SAFE maps across networks and draw meaningful conclusions about their differences and similarities.

In summary, SAFE provides a global perspective into the functional organization of a network by mapping statistical associations between functional groups and network regions. In contrast to most other methods for network analysis, which extract network regions and analyze them individually, SAFE shows that network layouts, coupled with robust enrichment analysis, are a valid strategy for studying intact molecular networks and gaining insight into the biological systems they represent.

EXPERIMENTAL PROCEDURES

SAFE is available at https://bitbucket.org/abarysh/safe and at http://dx.doi.org/10.17632/wdxwy8gmrz.1 (Baryshnikova, 2016), both as MATLAB code and a Mac OS X application that does not require MATLAB. Users are strongly encouraged to subscribe to the Bitbucket repository to receive notifications about code updates and bug fixes.

A step-by-step description of the algorithm, including key inputs and outputs, is provided in the Supplemental Experimental Procedures. Instructions for installation and several usage examples are provided in the README file of the Bitbucket repository.

Annotation of the Yeast GIS Network

The yeast genetic interaction similarity (GIS) network was constructed as described in Costanzo et al., (2010). Briefly, similarity of genetic interaction profiles for all pairs of 1,712 query genes and all pairs of 3,885 array genes were computed using Pearson correlation coefficients (R). Correlation values for gene pairs tested both as queries and as arrays were averaged. Gene pairs with similarities greater than $R = 0.2$ were connected in a network and visualized using the edge-weighted spring-embedded layout in Cytoscape (v.2.8) (Shannon et al., 2003), which implements the Kamada-Kawai force-directed algorithm (Kamada and Kawai, 1989). This algorithm positions each node based on the balance of attractive and repulsive forces exerted, respectively, by nodes connected and disconnected from it. Since the initial node positions are seeded randomly, the layout is non-deterministic and converges on different final positions at every run. Despite choosing the edge-weighted version of the algorithm, it was later discovered that a bug in Cytoscape (v.2.8) caused edge weights to be ignored. Thus, running the edge-weighted layout algorithm was equivalent to binarizing the network at the $r = 0.2$ threshold and running the unweighted version of the algorithm. The bug was fixed in Cytoscape (v.3.0).

The gene ontology (GO) biological process data and the yeast gene association files were downloaded from http://geneontology.org/ on August 19, 2014 (Ashburner et al., 2000). Annotations were propagated from child to parent terms, such that a gene was associated with a GO term if it was directly annotated to the term or any of its descendants.

The chemical genomics dataset was downloaded from the Dryad digital repository (Hoepfner et al., 2013) on December 3, 2013. Only homozygous profiling (HOP) data for 132 chemical compounds with known modes of action, as listed in the Table S1 of (Hoepfner et al., 2014), were used.

The quantitative genetic interaction data involving 13 members of the proteasome regulatory subunit were obtained and processed as described previously (Baryshnikova et al., 2010a, 2010b). Specifically, the laboratories of Charles Boone and Brenda Andrews (University of Toronto) conducted genome-wide synthetic genetic array (SGA) experiments to construct double mutants involving a deletion or a temperature-sensitive allele of a proteasome member (*RPN1*, *RPN5*, *RPN6*, *RPN7*, *RPN10*, *RPN11*, *RPN12*, *RPT1*, *RPT2*, *RPT3*, *RPT4*, *RPT6*, and *SEM1*) and the deletion mutants of all non-essential genes in the yeast genome. Quantitative double mutant fitness values were measured using colony size and compared to the fitness values of the two corresponding single mutants to produce a genetic interaction score that quantifies negative and positive deviations of the observed double mutant fitness from the expected combination of the two singles. The genetic interaction scores for these 13 SGA screens are available as Data S2.

Annotation of the Yeast Protein-Protein Interaction Network

A complete set of yeast protein-protein interactions was downloaded from BioGRID on April 26, 2015. The dataset was filtered to include only "physical" interaction types and exclude "biochemical activity" and "protein-RNA" experiments. The network was visualized using the spring-embedded layout in Cytoscape (v.3.2) (Shannon et al., 2003).

Source Data

The (1) SAFE source code implemented in MATLAB and sample data files and (2) the SAFE.app version 1.3 and sample data files that enable to run SAFE without MATLAB have been published in Mendeley Data and are available at http://dx.doi.org/10.17632/wdxwy8gmrz.1. Visit https://bitbucket.org/abarysh/safe for more information and updates.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three data files and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.014.

AUTHOR CONTRIBUTIONS

A.B. conceived the idea, performed the analysis, and wrote the manuscript.

ACKNOWLEDGMENTS

I am deeply grateful to Dmitriy Gorenshteyn for his invaluable help with writing this manuscript. Also, I would like to thank Amanda Amodeo, David Botstein, Michael Costanzo, Charles Boone, and Brenda Andrews for reading the manuscript and making useful suggestions. The C.B. and B.A. labs generously provided the proteasome genetic interaction data. This work was supported in its entirety by the Lewis-Sigler fellowship at Princeton University.

Received: October 11, 2015
Revised: February 9, 2016
Accepted: April 18, 2016
Published: May 26, 2016

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B., and Boone, C. (2010a). Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. Methods Enzymol. *470*, 145–179.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., et al. (2010b). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat. Methods *7*, 1017–1024.

Baryshnikova, A., Costanzo, M., Myers, C.L., Andrews, B., and Boone, C. (2013). Genetic interaction networks: toward an understanding of heritability. Annu. Rev. Genomics Hum. Genet. *14*, 111–133.

Botstein, D., and Fink, G.R. (2011). Yeast: an experimental organism for 21st century biology. Genetics *189*, 695–704.

Chen, D., Frezza, M., Schmitt, S., Kanwar, J., and Dou, Q.P. (2011). Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives. Curr. Cancer Drug Targets *11*, 239–253.

Cornish, A.J., and Markowetz, F. (2014). SANTA: quantifying the functional content of molecular networks. PLoS Comput. Biol. *10*, e1003808.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Cusick, M.E., Klitgord, N., Vidal, M., and Hill, D.E. (2005). Interactome: gateway into systems biology. Hum. Mol. Genet. *14* (Spec No. 2), R171–R181.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Hernández, F., and Cannon, M. (1982). Inhibition of protein synthesis in Saccharomyces cerevisiae by the 12,13-epoxytrichothecenes trichodermol, diacetoxyscirpenol and verrucarin A. Reversibility of the effects. J. Antibiot. *35*, 875–881.

Ho, C.H., Piotrowski, J., Dixon, S.J., Baryshnikova, A., Costanzo, M., and Boone, C. (2011). Combining functional genomics and chemical biology to identify targets of bioactive compounds. Curr. Opin. Chem. Biol. *15*, 66–78.

Hoepfner, D., Helliwell, S.B., Sadlish, H., Schuierer, S., Filipuzzi, I., Brachat, S., Bhullar, B., Plikat, U., Abraham, Y., Altorfer, M., et al. (2013). Data from: high-resolution chemical dissection of a model eukaryote reveals targets, pathways and gene functions. http://dx.doi.org/10.5061/dryad.v5m8v.

Hoepfner, D., Helliwell, S.B., Sadlish, H., Schuierer, S., Filipuzzi, I., Brachat, S., Bhullar, B., Plikat, U., Abraham, Y., Altorfer, M., et al. (2014). High-resolution chemical dissection of a model eukaryote reveals targets, pathways and gene functions. Microbiol. Res. *169*, 107–120.

Kamada, T., and Kawai, S. (1989). An algorithm for drawing general undirected graphs. Inf. Process. Lett. *31*, 7–15.

Khatri, P., Sirota, M., and Butte, A.J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. *8*, e1002375.

Kobourov, S.G. (2012). Spring embedders and force directed graph drawing algorithms. arXiv:12013011v1.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Lee, A.H., Iwakoshi, N.N., Anderson, K.C., and Glimcher, L.H. (2003). Proteasome inhibitors disrupt the unfolded protein response in myeloma cells. Proc. Natl. Acad. Sci. USA *100*, 9946–9951.

Lipson, C., Alalouf, G., Bajorek, M., Rabinovich, E., Atir-Lande, A., Glickman, M., and Bar-Nun, S. (2008). A proteasomal ATPase contributes to dislocation of endoplasmic reticulum-associated degradation (ERAD) substrates. J. Biol. Chem. *283*, 7166–7175.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabási, A.L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science *347*, 1257601.

Mitra, K., Carvunis, A.R., Ramesh, S.K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. Nat. Rev. Genet. *14*, 719–732.

Mo, M.L., Palsson, B.O., and Herrgård, M.J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst. Biol. *3*, 37.

Murray, M.Y., Auger, M.J., and Bowles, K.M. (2014). Overcoming bortezomib resistance in multiple myeloma. Biochem. Soc. Trans. *42*, 804–808.

Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C., and Troyanskaya, O.G. (2006). Finding function: evaluation methods for functional genomic data. BMC Genomics *7*, 187.

Newman, M.E. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA *103*, 8577–8582.

Obeng, E.A., Carlson, L.M., Gutman, D.M., Harrington, W.J., Jr., Lee, K.P., and Boise, L.H. (2006). Proteasome inhibitors induce a terminal unfolded protein response in multiple myeloma cells. Blood *107*, 4907–4916.

Parsons, A.B., Brost, R.L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G.W., Kane, P.M., Hughes, T.R., and Boone, C. (2004). Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat. Biotechnol. *22*, 62–69.

Schappert, K.T., and Khachatourians, G.G. (1986). Effects of T-2 toxin on induction of petite mutants and mitochondrial function in Saccharomyces cerevisiae. Curr. Genet. *10*, 671–676.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Sharan, R., and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. Nat. Biotechnol. *24*, 427–433.

Siniossoglou, S., Peak-Chew, S.Y., and Pelham, H.R. (2000). Ric1p and Rgp1p form a complex that catalyses nucleotide exchange on Ypt6p. EMBO J. *19*, 4885–4894.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Tacar, O., Sriamornsak, P., and Dass, C.R. (2013). Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. J. Pharm. Pharmacol. *65*, 157–170.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. Science *320*, 1465–1470.

Yu, C., Rahmani, M., Conrad, D., Subler, M., Dent, P., and Grant, S. (2003). The proteasome inhibitor bortezomib interacts synergistically with histone deacetylase inhibitors to induce apoptosis in Bcr/Abl+ cells sensitive and resistant to STI571. Blood *102*, 3765–3774.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104–110.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes Dev. *21*, 1010–1024.